Online learning for distributed optimal control of an electric vehicle fleet

Roman Le Goff Latimier, Guénolé Chérot, H. Ben Ahmed

Online Learning for Distributed Optimal Control of an Electric Vehicle Fleet

R. Le Goff Latimier, G. Chérot, H. Ben Ahmed
SATIE, ENS Rennes, CNRS, Bruz, France
{roman.legoff-latimier, guenole.cherot, benahmed}@ens-rennes.fr

***Abstract*—The management of electrical power systems requires the resolution of large-scale problems whose agents are linked by coupling constraints. Nevertheless, decomposition methods cannot provide an exact solution while dealing with temporal dynamics in a stochastic environment. Each agent would then have to solve a local problem in which future quantities intervene. However they depends on other agents' future decisions which are still unknown. In order to enhance the existing approximate approaches to this problem, the proposed method involves Alternating Direction Method of Multipliers to overcome the large dimension by an iterative resolution of local coordinated problems. Uncertain temporal dynamics are handled by a stochastic dynamic programming approach. In order to make local problems tractable, an online learning step is added. The agents can then anticipate future global variations in a local but uncertain way. The optimal charging of an electric vehicle fleet paired with a wind power plant is considered as a case study. The expected benefits are highlighted, both at the outset and after training the anticipatory models. The discussion addresses the learning parameters allowing the fastest convergence.**

***Index Terms*—Alternating Direction Method of Multipliers, Decomposition Method, Stochastic Dynamic Programming, Virtual Power Plant, Online Learning**

I. Introduction

The current deployment of distributed energy resources on power systems introduces several complexity sources in their management problems. First, the ongoing installation of numerous electrochemical storage units induces new temporal dynamics. This relates to stationary batteries of large capacity, but also domestic batteries scattered on distribution networks as well as electric vehicles [1]. Controllable loads have also long contributed to this time coupling through heating and domestic hot water. By definition, these storages create temporal dynamics on energy exchanges that should be taken into account, in particular at the distribution level [2], [3]. It is therefore necessary to anticipate behaviors, although this exercise is facing increasing difficulties. On the one hand, decentralized productions, mainly intermittent renewables such as photovoltaic and wind power, add local effects to the problem because of their meteorological sensitivity [4]. On the other hand, forecasting becomes increasingly difficult when it comes to very small scales, such as within distribution districts. A further aspect of this situation is the multiplication of involved actors. Due to the simultaneous emergence of decentralized production, information technology and new usages such as the electric vehicle, more and more consumers are becoming prosumers, seeking to control their consumption and even their production [5].

These complexities raise challenges not only for operational management methods, but also for the approaches proposed in the literature. It is indeed commonly found to overcome one or another of the difficulties like large scale, stochasticity or time coupling. But dealing simultaneously with all of them is much more difficult. For example, market mechanisms are extremely efficient for managing large-scale problems, while they transfer the issues of uncertainty and temporal coupling to the stakeholders [6], [7]. The simultaneous handling of these different complexities has given rise to a very rich and dynamic literature. Numerous approaches have been proposed, albeit without achieving an exact resolution of the problem. Moreover, among the existing approximate methods, the comparative analyses permitting to identify the most adapted ones are still largely to be realized. These numerous works recurrently rely on dual decompositions [8], [9], Markov Decision Process [10] or on learning methods [11]–[14].

The present contribution intends to propose an original method for solving large scale problems with coupling constraints, temporal dynamics and stochastic components. It is based on an Alternating Direction Method of Multipliers ADMM decomposition [15] to subdivide the initial problem into local problems coordinated with each other. These would then fall within the scope of Stochastic Dynamic Programming [16]. However, the exact resolution of the local problems is currently impossible as future global costs cannot be known ar the local scale. Taking advantage from multiagents learning approaches, we then propose to add an online learning step so that each agent iteratively builds a forecast for the evolution of these unknown quantities.

The continuation of this article is structured as follows. In section II, a representative case study will be presented. Within this illustrative context, the proposed method will be described in the section III and compared with the most similar approaches. The main results will be presented in section IV, as well as a discussion of the parameters that need to be adjusted in order to make the most of the proposed decomposition. Finally, section V will summarise the main ideas of this contribution as well as the perspectives for future work.

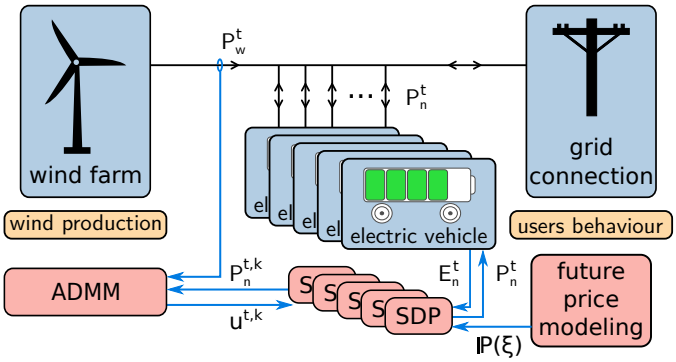

Fig. 1. Illustrative test case under consideration: a fleet of electric vehicles maximising its recharge with energy from a wind power plant. Blue: physical actors; orange: stochastic elements; red: optimal control architecture.

II. Description of the illustrative case

In order to illustrate the proposed method, this section describes a representative case study of high-dimensional stochastic problems involving temporal dynamics and coupling constraints [10], [17].

As described in figure 1, it consists in a fleet of 200 electric vehicles, combined with a wind power plant. This virtual power plant's purpose is to maximize the recharge using directly the wind power produced. Wind generation data is a 3-year time series from the Bonneville Power Administration open source dataset[1]. The peak power is scaled according to the different studies presented in the section IV. The considered time step ΔT is hourly. The vehicles are assumed to make two commutes per day, whose arrival and departure times are randomly selected. For this purpose, French national statistics are employed [18]. They describe the departure and arrival times between residence and workplace. These statistics also provide the distribution of the travelled distances, which can therefore be randomly drawn. This distance is converted into energy expended during the trip, allowing to deduce the initial state of charge of the vehicle when it is connected to a charging station. The capacities of the various electric vehicles in the fleet are also randomly selected according to the market share of electric vehicles in Europe. For each travel – which is unique because randomly drawn – the user is supposed to indicate a departure time and the minimum energy level he expects by then – again randomly drawn. In order to allow the reproduction of the presented results and to propose a benchmark problem on which to compare resolution methods later on, all the data are available on the following gitlab repository: https://gitlab.com/satie.sete/online-learning-for-distributed-optimal-control.

A control problem is proposed for the virtual plant, where $\mathcal{N}$ stands for the set of vehicules. For the sake of illustration, it is designed to be as simple as possible. Two situations are penalized. First one is if electric vehicle fleet consumes more than the instantaneous plant's production P_w^t. Power must therefore be pulled from the grid. Second one is if, at its departure time t_n^{dep}, a vehicle is not charged to at least the level specified by the user E_n^*.

These two penalties being homogeneous to energies, they are gathered in the following problem, with $\lfloor\cdot\rfloor_+ := \max(0,\cdot)$:

$$\min_{(P_n^t)} \mathbb{E}\Big(\sum_t \Delta T \left\lfloor -P_w^t + \sum_n P_n^t \right\rfloor_+ + \sum_{n\in\mathcal{N}} \mathbb{1}_{t=t_n^{\text{dep}}} \left\lfloor E_n^* - E_n^t \right\rfloor_+ \Big) \tag{1a}$$

s.t. $\forall t, \forall n$,

$$E_n^{t+\Delta T} = E_n^t + \Delta T \cdot P_n^t \tag{1b}$$

$$0 \leqslant E_n^t \leqslant E_n^\sharp \tag{1c}$$

$$P_n^\flat \leqslant P_n^t \leqslant P_n^\sharp \tag{1d}$$

in which P_n^t is the charging power of vehicule n at time t – $P_n^t > 0$ for charging – P_w^t the instantaneous wind plant production and E_n^* the energy level wished by the user n upon his departure time t_n^{dep}. $E_n^\sharp$ represents the battery capacity of the vehicle n. $P^{\flat,\sharp}$ denotes the minimum and maximum powers for the vehicle n. In accordance with the performance of current high-power charging stations, they are set to allow for a full recharge in 1 h. In (1), the expectation is to be evaluated against several stochastic quantities. The future wind generation is unknown, as well as the behavior of the vehicles that will connect after the considered time.

III. Proposed method

This section presents the proposed method from the case study described in section II. Problem (1) can be decomposed using ADMM by introducing the variable $\overline{Q}$ subject to the constraint $\overline{Q} = \overline{P}$, where $\overline{P}$ denotes the average power of the vehicles. The following iterations can then be deduced, with k the current iteration:

- $\forall n,\ P_n^{t,k+1} := \underset{P_n^t}{\text{argmin}}\ \mathbb{1}_{t=t_n^{\text{dep}}} \left\lfloor E_n^* - E_n^t \right\rfloor_+ + \frac{\rho}{2}\Big(P_n^t - P_n^{t,k} + \overline{P}^{t,k} - \overline{Q}^{t,k} + u^{t,k}\Big)^2 +$

$$\mathbb{E}\Big(\frac{\rho}{2}\sum_{\tau=t+\Delta T}^{t_n^{\text{dep}}} \big(P_n^\tau - P_n^{\tau,L} + \overline{P}^{\tau,L} - \overline{Q}^{\tau,L} + u^{\tau,L}\big)\Big) \tag{2a}$$

$$\text{s.t. } E_n^{t+\Delta T} = E_n^t + \Delta T \cdot P_n^t$$
$$0 \leqslant E_n^t \leqslant E_n^\sharp$$
$$P_n^\flat \leqslant P_n^t \leqslant P_n^\sharp$$

- $\overline{Q}^{t,k+1} := \underset{\overline{Q}^t}{\text{argmin}} \left\lfloor N\overline{Q}^t - P_w^t \right\rfloor_+ + \frac{N\rho}{2}\big(\overline{Q}^t - \overline{P}^{t,k+1} - u^{t,k}\big)^2$ (2b)

- $u^{t,k+1} := u^{t,k} + \overline{P}^{t,k+1} - \overline{Q}^{t,k+1}$ (2c)

The first update (2a) deals with the local problem for each agent. This problem has several terms, here separated by line breaks. The first line is about the individual and known

[1] https://transmission.bpa.gov/business/operations/Wind/twndbspt.aspx

objective of the vehicle: its own recharge for its departure time. The second line gathers the coordination terms from the ADMM decomposition. These terms ensure that the decisions of all vehicles will converge to a consensus decision at the end of the k iterations. They involve the individual vehicle power at the previous iteration $P_n^{t,k}$, the average recharge at the previous iteration $\overline{P}^{t,k}$ and its corresponding variable $\overline{Q}^{t,k}$, as well as the scaled dual variable $u^{t,k}$. Finally the third line concerns the anticipation that the vehicle must have of future coordination costs. They cannot be known for sure. For every future time period, the same terms from the ADMM will therefore intervene. Note that the iterations on k are only transient steps of the current resolution. When a vehicle seeks to anticipate future coordination costs, only the values at convergence will be important. L then stands for the final iteration of ADMM algorithm at future time steps. As convergence will then be reached, $\overline{Q}^{\tau,L} = \overline{P}^{\tau,L}$ allows to simplify these terms. The dynamic and boundary constraints on energy and power apply individually to each of these local problems.

The step (2b) performs the update of the global charging problem of the fleet as a whole. N refers to the number of electric vehicles connected at this given time step. This update has no dynamic constraint since it has no direct control over the charging powers. Only the average power of the vehicles at iteration $k+1$, $\overline{P}^{t,k+1}$ is involved. Finally the step (2c) updates the scaled dual variable, so as to enforce the convergence of $\overline{Q}^t = \overline{P}^t$. These steps must be iterated until convergence of the primal and dual residues.

However, the resolution of the (2a) step is not straightforward because at this stage a single vehicle has no element to minimize the expectation term of the future consensus costs: $P_n^\tau - P_n^{\tau,L} + u^{\tau,L}$. This term depends on the behavior of the other users, on the wind production, on the future situation of the fleet as a whole. The approach proposed here begins with considering these unknown terms as a random variable, noted

$$\xi_n^\tau = P_n^{\tau,L} - u^{\tau,L} \tag{3}$$

The challenge then becomes to establish a model for the evolution of this new variable $\mathbb{P}\left(\xi_n^{\tau+\Delta T} | \xi_n^\tau\right)$. The presented methodology then builds it iteratively on the basis of the observations that will be made as the exercise is going on. The first initialization can be done with a rudimentary model, for example a persistence or a null value at the next time step.

Assuming that the evolution of ξ_n is described by a probabilistic model – even if this one is not faithful to reality, in particular at the beginning of the training – it becomes then possible to apply a complete resolution to the problem (2a). Several methods could be considered, such as a model predictive control or a decomposition by scenarios. The retained option here is Stochastic Dynamic Programming SDP [16]. Indeed, many vehicles must simultaneously solve the same problem, each one being in a different situation but described by the same quantities. SDP allows to compute a unique optimal strategy, describing the optimal solution for any value of a state vector which is here composed of:

$$x_n^t = \left(E_n^\sharp, E_n^*, E_n^t, \xi_n^t\right) \tag{4}$$

The control vector consists solely of the charging power P_n^t. Since this state vector is focused on a vehicle local problem, the time t used so far to set the global problem in the meaning of the hour of the day is not the most relevant. Indeed the individual problem of each vehicle will come to an end when the vehicle leaves at t_n^{dep}. We therefore introduce a remaining recharge time $d_n = t_n^{\text{dep}} - t$, which will be the time used during the resolution of the local vehicle problem by SDP. The Bellman equation can then be applied to computed the value function V on a grid discretizing the state space $\mathbb{X}$:

$$\bullet \quad \forall x \in \mathbb{X}, \;\; V\big(d=0, x\big) = \lfloor E_n^* - E_n \rfloor_+ \tag{5a}$$

$$\bullet \quad \forall d > 0, \forall x \in \mathbb{X}, \;\; V\big(d, x\big) = \min_{P_n} \frac{\rho}{2}\big(P_n - \xi_n^d\big)^2 + \mathbb{P}\left(\xi_n^{d-\Delta T} | \xi_n^d\right) \cdot V\big(d - \Delta T, x^{d-\Delta T}\big) \tag{5b}$$

Within the state at the future epoch $x^{d-\Delta T}$, the terms $E_n^\sharp$ and E_n^* are constant. The term E_n^t evolves deterministically according to the dynamic equation (1b). Only the term ξ is stochastic and must therefore be weighted by its probability of occurrence $\mathbb{P}\left(\xi_n^{d-\Delta T} | \xi_n^d\right)$. Once the problem becomes solvable, the remaining task is to learn the model iteratively. But updating it alters the actions of the vehicles, and therefore modifies the ξ dynamic to be learned. Hence all vehicles cannot be updated simultaneously as this would likely create instabilities. In consequence, the fleet is split into subgroups $\mathcal{N}_i$ that are to be updated one after the other.

The algorithm here proposed can be related to several features of the literature. First, extensions of dynamic programming propose several approaches to overcome the curse of dimensionality. Linear systems allow formulations such as the Stochastic Dual Dynamic Programming [19], [20]. In the convex case, the introduction of approximation is necessary [21], similarly to our approach. In particular, the method known as Dual Approximate Dynamic Programming [22] uses a dual decomposition rather than an ADMM. In addition, the evolution of the dual variable of the problem is there described by an exogenous model, typically autoregressive, whose coefficients are then to be learned. Besides, multi-agent reinforcement learning shares the procedure of updating the strategy according to the accumulated observations. The parameters of the updates can therefore be discussed in the light of this literature, in order to improve the speed of convergence while avoiding possible oscillations. Moreover, the question of the value of the shared observation is common between this field and the present approach.

IV. Results and discussion

In the following section, the method proposed in section III is applied on the illustrative case presented in section II, with $\rho = 1$. The individual control of each vehicle is realized

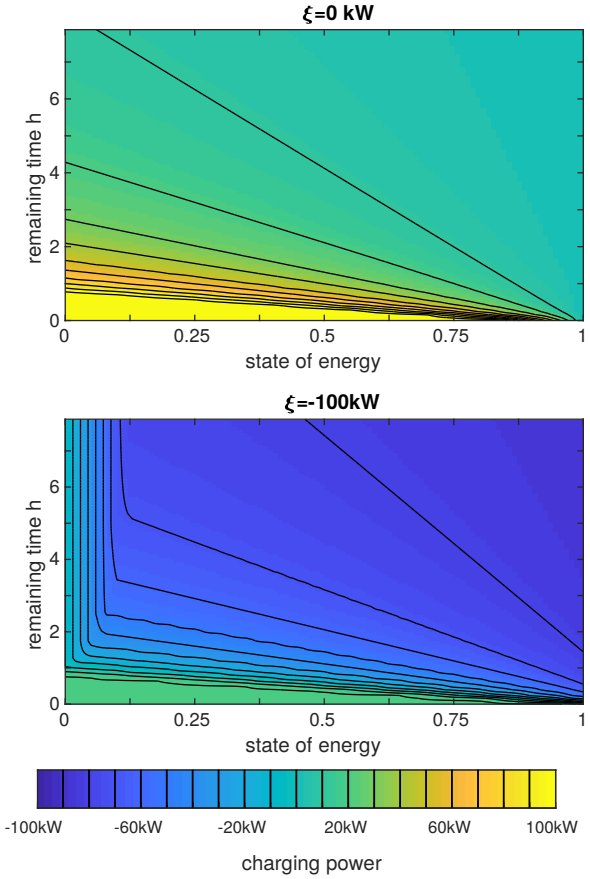

Fig. 2. Cross-sections of the optimal charging strategy along the plane (energy state, remaining charging time) for two ξ values. The isopowers lines are plotted every 10kW. $E_n^\sharp = 85\,\text{kWh}$, $E_n^* = 85\,\text{kWh}$

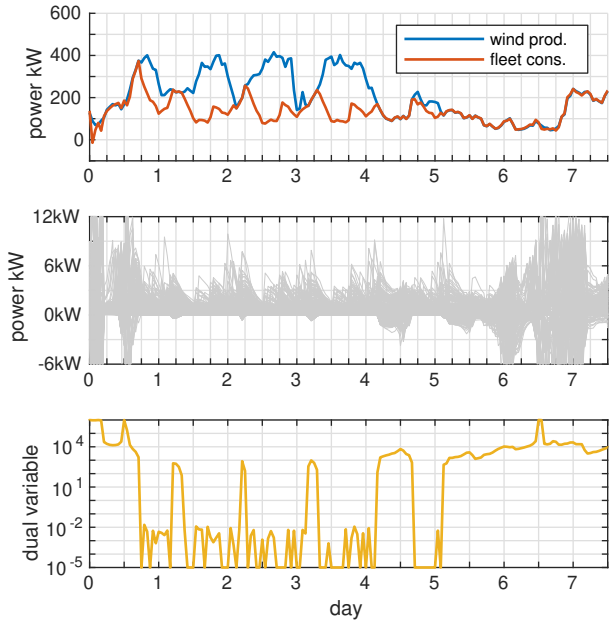

Fig. 3. Sample of the optimal charging times serie for the vehicle fleet. Top: power produced by the wind power plant and total charging power of the fleet. Middle: individual powers of each vehicle. Bottom: dual variable of the sharing problem (logarithmic scale). $P_w^* = 1MW$.

using the optimal strategy resulting from solving the Bellman equation (5). This strategy thus describes the optimal charging power to be applied for any configuration of the state vector. Figure 2 represents cross-sections of this strategy within the plane (E_n, d_n) – stored energy versus time remaining before departure – for a battery capacity of $85\,\text{kWh}$ and a will of a fully recharged battery. For $\xi = 0$, the response of this strategy is natural: vehicles that have limited recharge time left and are still lightly charged must recharge substantially – lower left corner – within their maximum recharge power $P_n^\sharp = 85\,\text{kW}$. On the contrary, vehicles leaving in a long time and already with important stored energy apply powers close to 0 – upper right corner. When the coordination variable ξ becomes negative, this strategy is shifted to negative powers: vehicles are incentivized to discharge. A gradient is also noticeable, as vehicles with more time ahead of them discharge sooner than the more constrained ones. If the coordination variable ξ takes even smaller values, all vehicles can be compelled to unload, whatever their situation.

The implementation of this control based on ADMM and SDP is illustrated in Figure 3 for a 200 vehicle fleet and a 1MW plant. Convergence is supposed reached when residues fall below $100\,\text{W}$. The first panel displays the wind power produced, as well as the total power of the fleet. The regulation to not consume more than the wind generation is manifested on days 1 and 4 to 8. The trajectories of the individual charging powers for each vehicle are visible in the second panel. In particular it is noticeable that some vehicles discharge punctually when the wind power is low. Some vehicles with ample time left inject power in favor of other more urgent vehicles. The consensus to reduce costs at the fleet level is thus highlighted. Finally, the evolution of the dual variable is described on the third panel in logarithmic scale. Abrupt fluctuations can be observed whenever the wind power produced is not high enough.

For the sake of illustration, the time series samples shown in figure 3 were obtained with an optimal strategy calculated before training the transition probability model $\mathbb{P}\left(\xi(t + \Delta T)|\xi(t)\right)$. The basic model used for initialization was $\mathbb{P}\left(\xi(t + \Delta T) = 0\right) = 1$, which is close to an approximation of future consensus costs at zero. In spite of this naïve anticipation, we can notice that the behaviors generated by the proposed method are consistent. Indeed, the proposed method allows to exactly take into account the present and future individual costs, as well as the consensus costs at the present time. Only the future consensus costs need to be progressively learned. Hence, the performances are already satisfying from the initialization.

The application of this charging strategy allows to generate data on the variation of the coordination variable ξ. It then becomes possible to develop the probability matrix of this variable, illustrated in figure 4 for two wind power plant capacities.

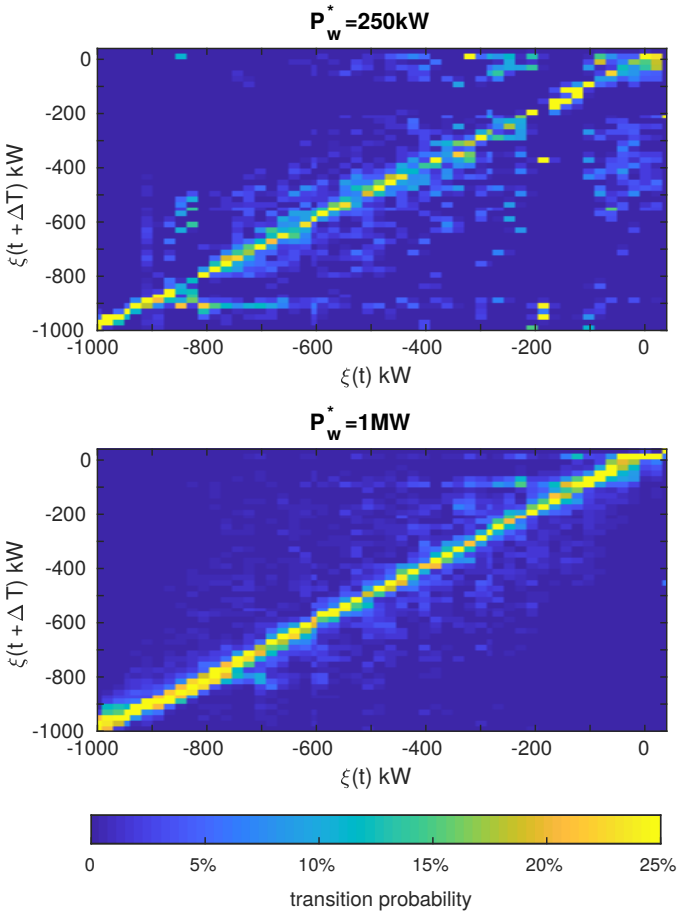

Fig. 4. Transition probabilities $\mathbb{P}\left(\xi(t+\Delta T)|\xi(t)\right)$ identified for two capacities of wind plant.

It is remarkable that for a high capacity plant, situations of production too low for the fleet are rarer – and more perennial. Thus, the ξ variations have slow dynamics, hence an evolution very close to a simple persistence. Conversely, an undersized plant causes frequent shortages of shorter duration. This results in far more erratic variations in the ξ variable which cannot be anticipated intuitively. Consequently, the performance of the proposed method must be assessed with respect to this dependence of the model to be learned on the capacity of the plant.

The costs associated with the problem (1) are described in figure 5 as a function of the capacity of the wind power plant. A large capacity leads to very low costs: all vehicles easily find wind power to charge. Using an online update of the strategy, the observed gains can therefore only be marginal. The evolution model to be identified is very simple. Conversely, low capacities cause significantly higher overall costs because vehicles often have to charge using non-wind electricity. The use of online learning is then followed by important absolute reductions because the evolution model is difficult to identify and brings an important added value. However, it is notable that relative reductions (visible on the bottom panel) are not maximal for similar situations but have

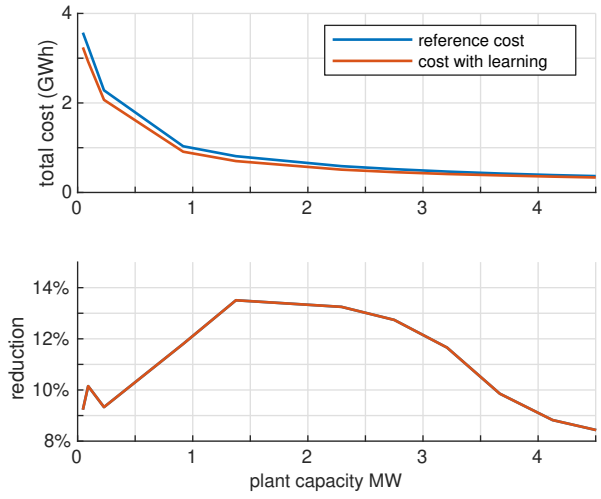

Fig. 5. Top: Evolution of the operating cost as a function of the capacity of the wind power plant over a 3-year simulation. The reference cost in blue does not use observations to update its management strategy, unlike the cost in red. Bottom: normalized reduction.

a peak value for a wind capacity around 2 MW. The fleet nominal annual consumption being 1 GWh/year, this translates into an annual production which is four to five times higher than the needs, based on the dataset's producible power.

Many parameters can affect the convergence speed of online learning. Some of them are illustrated in figure 6 which shows the performances obtained by strategies sampled at various stage of training, on a one year test series – different from the one used for the learning. The reference case in blue consists of a learning process where 10% of the vehicles update their strategy every day, using the observations made by the whole fleet. After the first 30 days, convergence seems

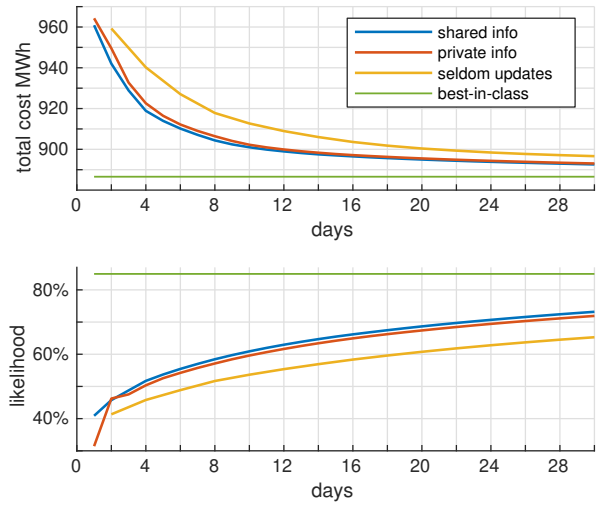

Fig. 6. Evolution of the strategy performance according to the online learning method, computed over a 1-year time series. $P_w^* = 150kW$

to be established. The performance obtained by a strategy using 1000 days of training is however indicated in green. It is thus noticeable that the training will continue to improve its performance, albeit more and more slowly. The first variation shown relates to the frequency of updates of the charging strategy – curve in yellow. The vehicles are then updated only every two days, thus based on more observations. The convergence speed is consequently divided by two. This indicates that the accumulation of observation data generated by a previous strategy does not lead to a better update. It is useless to learn behaviors caused by poorly estimated strategies. Instead, updating as frequently as possible is better. This result is similar to the behaviors highlighted in multi-agent learning [23]. Additionally, the red curve reflects the value of the shared information. The reference situation is here modified by updating the strategies only on the basis of the observations made by the updated vehicles. In this situation, the fleet is divided into 10 groups that share their information only among themselves and are updated simultaneously. The result – counterintuitive – is that the speed of convergence is indeed reduced, but marginally. This can be explained considering that the variations of ξ are mainly determined by the dual variable u which is unique for the whole fleet. Pooling the observations of all the vehicles therefore only allows to mutualize situations that are very close to each other. The lower panel of figure 6 illustrates the variations of the likelihood of the ξ model: how well has the transition matrix anticipated the observed variations. It is remarkable that the convergence on the costs appears much faster than the convergence on the likelihood. Models with low fidelity thus allow for a rapid improvement in the performance obtained. Furthermore, the configuration where information is not universally shared leads to strategies that are based on subsets of the observations, without any guarantee of continuous improvement. The non-regular evolution of the strategy with private information can be observed between days 2 and 3.

In order to further study the speed of convergence and the impact of information sharing, the number of vehicles updated simultaneously is modified according to the values indicated in the table IV. The duration is then sought before the resulting strategy produces operating costs close to those of the strategy trained on $1000,\text{days}$. An extra cost of 1% is adopted as a convergence criterion. Initially, the information is shared among all vehicles and the number of vehicles updated daily is modified. It can be seen that this leads to a slight increase in the convergence time. Besides, for a ratio of $50/50$ of vehicles updated and sharing all their observations, oscillations appear on the performances of the strategies: they do not improve

	10%	20%	50%
shared info	19 d	20 d	21 d
private info	21 d	20 d	21 d

TABLE I
TIME UNTIL STRATEGY CONVERGENCE, IN DAYS

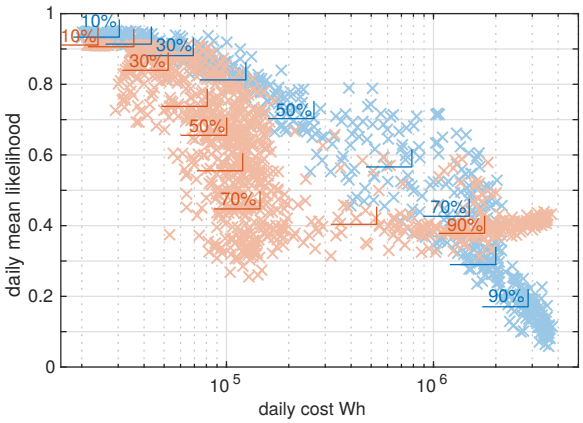

Fig. 7. Daily costs depending on the daily average likelihood over 1000 day time series, according to a naive strategy in blue and a trained strategy in red. Markers: deciles in cost and likelihood. $P_w^* = 2\,\text{MW}$.

monotonously anymore. This can be imputed to the fact that too many agents change their strategies simultaneously and that the dynamics of the dual variable is no longer faithful to previous observations. This encourages updating as few agents as possible to improve both the stability and the speed of convergence. This observation is again in line with the practices of multi-agent learning [23]. In a second step, the impact of information sharing is investigated. We can notice that it can cause a slowing down of the convergence speed, which is expected. However, this slowdown is less and less noticeable when the subgroups get larger and larger.

The last analysis here proposed regards the link between the likelihood and the evolution of the cost, as illustrated in figure 7. For each day of a 1000 day time series, the daily cost is represented depending on the average likelihood on that day. Two strategies are compared: in blue a naive initialization strategy and in red a trained strategy. The deciles in cost and likelihood are shown for both scatter plots: 10% of the days are to the left and above the first marker. We can see that a trained strategy will not succeed in reducing the extreme costs: these are days with no wind generation, so the best charging strategy would not make any difference. However, these days are better explained. The vast majority of the points illustrate that the trained strategy reduces the costs with the same likelihood. Finally, the two strategies have similar performances on very windy days where the production is sufficient to recharge the fleet anyway.

V. CONCLUSION AND PERSPECTIVES

This contribution addressed large scale problems under a coupling constraint and involving stochastic time dynamics. A representative case study has been presented and documented, a fleet of electric vehicles aiming at maximizing its recharge by a wind power plant. The proposed solution is based on Alternating Direction Method of Multipliers and Stochastic Dynamic Programming. In a first step, the global large-scale problem is decomposed according to ADMM. Local problems

with temporal dynamics and stochastic behavior could then be solved by SDP. However, local problems do not know the future evolution of the dual variable associated with the problem. It is therefore proposed that they progressively build a probabilistic model of its evolutions. The results presented have highlighted the performances obtained by this method. In particular, the learning dynamics have been illustrated, with a particular attention to the impact of information sharing between agents.

Several perspectives are raised by this study. First, the multiplicity of proposed management methods – to which this article is contributing – requires the adoption of common, documented and accessible benchmark problems. In order to enhance comparative analysis, the case study used is freely available, as well as the code of the proposed method[2]. This test case should be further enhanced by adding a grid topology to cover congestion management.

The addition of covariables within the transition probabilities would be an promising way to improve the likelihood of the transition model. As this likelihood is strongly linked to the performance of the control strategies, a significant improvement could be expected. However, such an sophistication of the transition model would necessarily require a much larger number of observations, increasing by an exponent equal to the number of co-variables involved. The learning process would therefore be slowed down, which could counterbalance the resulting overall improvement.

Moreover, the learning parameters could not here be investigated in an exhaustive manner. The sequence of observations and updates is essential to quickly converge on a successful strategy. However, as is well known in the field of multi-agent learning, some modalities could generate instabilities and never converge. For example, updating all agents at once leads to such instabilities. The instability conditions of the proposed approach still need to be investigated, as well as a universal method to determine the optimal parameters.

Finally, the last perspective of research that we will mention concerns the link between the performance obtained and the access to other vehicles' data. Only the cases of subgroups data and completely shared data have been considered here. The study presented could be supplemented by a data market where each vehicle could offer its observations. The purchase of observations would then be a bet on the expected improvement. The fairest mechanism would then consist of paying each observation according to the improvements it will have allowed on the strategies of the other vehicles.

REFERENCES

[1] R. Le Goff Latimier, B. Multon, and H. Ben Ahmed, "Distributed optimisation with restricted exchanges of information: Charging of en electric vehicle fleet," in *CIRED Workshop*, 2018.

[2] M. Di Somma, G. Graditi, E. Heydarian-Forushani, M. Shafie-khah, and P. Siano, "Stochastic optimal scheduling of distributed energy resources with renewables considering economic and environmental aspects," *Renewable energy*, vol. 116, pp. 272–287, 2018.

[2]https://gitlab.com/satie.sete/online-learning-for-distributed-optimal-control

[3] M. A. Gilani, A. Kazemi, and M. Ghasemi, "Distribution system resilience enhancement by microgrid formation considering distributed energy resources," *Energy*, vol. 191, p. 116442, 2020.

[4] D. Dongol, T. Feldmann, M. Schmidt, and E. Bollin, "A model predictive control based peak shaving application of battery for a household with photovoltaic system in a rural distribution grid," *Sustainable Energy, Grids and Networks*, vol. 16, pp. 1–13, 2018.

[5] P. Kotler, "The prosumer movement," in *Prosumer revisited*. Springer, 2010, pp. 51–60.

[6] Y. Zhou, A. Scheller-Wolf, N. Secomandi, and S. Smith, "Managing wind-based electricity generation in the presence of storage and transmission capacity," *Production and Operations Management*, vol. 28, no. 4, pp. 970–989, 2019.

[7] J. L. Crespo-Vazquez, T. AlSkaif, Á. M. González-Rueda, and M. Gibescu, "A community-based energy market design using decentralized decision-making under uncertainty," *IEEE Transactions on Smart Grid*, vol. 12, no. 2, pp. 1782–1793, 2020.

[8] G. Tsaousoglou, K. Mitropoulou, K. Steriotis, N. G. Paterakis, P. Pinson, and E. Varvarigos, "Managing distributed flexibility under uncertainty by combining deep learning with duality," *IEEE Transactions on Sustainable Energy*, 2021.

[9] K. De Craemer, S. Vandael, B. Claessens, and G. Deconinck, "An event-driven dual coordination mechanism for demand side management of phevs," *IEEE Transactions on Smart Grid*, vol. 5, no. 2, pp. 751–760, 2013.

[10] Y. Yang, Q.-S. Jia, G. Deconinck, X. Guan, Z. Qiu, and Z. Hu, "Distributed coordination of ev charging with renewable energy in a microgrid of buildings," *IEEE Transactions on Smart Grid*, vol. 9, no. 6, pp. 6253–6264, 2017.

[11] H.-F. Xu, Q. Ling, and A. Ribeiro, "Online learning over a decentralized network through admm," *Journal of the Operations Research Society of China*, vol. 3, no. 4, pp. 537–562, 2015.

[12] V. Lakshminarayanan, V. G. S. Chemudupati, S. K. Pramanick, and K. Rajashekara, "Real-time optimal energy management controller for electric vehicle integration in workplace microgrid," *IEEE Transactions on Transportation Electrification*, vol. 5, no. 1, pp. 174–185, 2018.

[13] K. L. López, C. Gagné, and M.-A. Gardner, "Demand-side management using deep learning for smart charging of electric vehicles," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 2683–2691, 2018.

[14] M. Shin, D.-H. Choi, and J. Kim, "Cooperative management for pv/ess-enabled electric vehicle charging stations: A multiagent deep reinforcement learning approach," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 5, pp. 3493–3503, 2019.

[15] S. Boyd, N. Parikh, and E. Chu, *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc, 2011.

[16] D. P. Bertsekas, *Dynamic programming and optimal control*. Athena scientific Belmont, 2000.

[17] R. Le Goff Latimier, B. Multon, H. Ben Ahmed, F. Baraer, and M. Acquitter, "Stochastic optimization of an electric vehicle fleet charging with uncertain photovoltaic production," in *2015 International Conference on Renewable Energy Research and Applications (ICRERA)*. IEEE, 2015, pp. 721–726.

[18] Commissariat Général au Développement Durable, "Electric vehicles in perspective, cost benefit analysis and potential demand (in french: Les véhicules électriques en perspective, Analyse coûts-avantages et demande potentielle)," 2011.

[19] A. Shapiro, "Analysis of stochastic dual dynamic programming method," *European Journal of Operational Research*, vol. 209, no. 1, pp. 63–72, 2011.

[20] A. Papavasiliou, Y. Mou, L. Cambier, and D. Scieur, "Application of stochastic dual dynamic programming to the real-time dispatch of storage under renewable supply uncertainty," *IEEE Transactions on Sustainable Energy*, vol. 9, no. 2, pp. 547–558, 2017.

[21] Z. Pan, T. Yu, J. Li, K. Qu, L. Chen, B. Yang, and W. Guo, "Stochastic transactive control for electric vehicle aggregators coordination: A decentralized approximate dynamic programming approach," *IEEE Transactions on Smart Grid*, vol. 11, no. 5, pp. 4261–4277, 2020.

[22] P. Carpentier, J.-P. Chancelier, V. Leclère, and F. Pacaud, "Stochastic decomposition applied to large-scale hydro valleys management," *European Journal of Operational Research*, vol. 270, no. 3, pp. 1086–1098, 2018.

[23] K. Zhang, Z. Yang, and T. Başar, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," *Handbook of Reinforcement Learning and Control*, pp. 321–384, 2021.